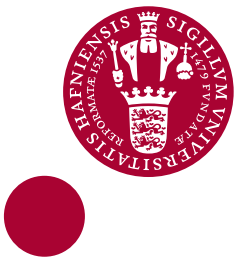

Raw material quality assessment approaches comparison in pectin production

André Fernandes Caroço, Ricardo; Bevilacqua, Marta; Armagan, Ibrahim; Santacoloma, Paloma A.; Abildskov, Jens; Skov, Thomas; Huusom, Jakob Kjøbsted

Published in:
Biotechnology Progress

DOI:
10.1002/btpr.2762

Publication date:
2019

Document version
Peer reviewed version

Document license:
Other

Citation for published version (APA):
André Fernandes Caroço, R., Bevilacqua, M., Armagan, I., Santacoloma, P. A., Abildskov, J., Skov, T., & Huusom, J. K. (2019). Raw material quality assessment approaches comparison in pectin production. *Biotechnology Progress*, *35*(2), 1-13. [e2762]. https://doi.org/10.1002/btpr.2762

Biotechnology Progress
DOI 10.1002/btpr.2762

Raw material quality assessment approaches comparison in pectin production

Ricardo F. Caroço[*§], Marta Bevilacqua[†], Ibrahim Armagan [‡]
Paloma A. Santacoloma[‡], Jens Abildskov[*], Thomas Skov[†], Jakob K. Huusom[*§]

October 30, 2018

Abstract

This article explores different opportunities to evaluate quality variation in raw materials from biological origin. Assessment of raw materials attributes is an important step in a bio-based production since fluctuations in quality are a major source of process disturbance. This can be due to a variety of biological, seasonal and supply scarcity reasons. The final properties of a product are invariably linked with the initial properties of the raw material. Thus, the operational conditions of a process can be tuned to drive the product to the required specification based on the quality assessment of the raw material being processed. Process analytical technology (PAT) tools which enable this assessment in a far more informative and rapid manner than current industrial practices that rely on rule-of-thumb decisions are assessed. An example with citrus peels is used to demonstrate the conceptual and performance differences of distinct quality assessment approaches. The analysis demonstrates the advantage of characterization through multivariate data analysis coupled with a complementary spectroscopic technique, near-infrared spectroscopy. The quantitative comparative analysis of three different approaches, discriminant classification based on expert-knowledge, unsupervised classification, and spectroscopic correlation with reference physicochemical variables, is performed in the same dataset context.

Keywords: Raw Material, Process Analytical Technology, Near-infrared spectroscopy, Chemometrics, Pectin.

[*]Process and Systems Engineering Centre (PROSYS), Department of Chemical and Biochemical Engineering, Technical University of Denmark, Søltofts Plads Building 229, DK-2800 Kgs. Lyngby, Denmark
[†]Chemometrics and Analytical Technology, Department of Food Science, University of Copenhagen, DK-1958 Frederiksberg, Denmark
[‡]CP Kelco ApS., Ved Banen 16, DK-4623 Lille Skensved, Denmark
[§]corresponding author: jkh@kt.dtu.dk

doi: 10.1002/btpr.2762

Received: Jun 08, 2018; Revised: Oct 30, 2018; Accepted: Nov 29, 2018

Introduction

Current traditional bioprocessing plant manufacturing does not tackle raw material variability with enough operational flexibility, resulting in product quality variability. Unlike other process conditions, the manufacturer does not directly control raw materials. The quality is highly dependent on external vendors. Furthermore, raw materials may vary from lot to lot on a long timescale. In these cases, the measurement of critical properties of the raw material can allow for the dynamic monitoring and control.[1,2] The need to tackle the natural variability of raw materials is considered an important challenge in industrial biotech processes. An adequate approach to this problem can mitigate production performance issues and undesired deviation of the critical quality attributes of the end-product due to raw material quality fluctuation.[3–5] Identification and measurement of key raw material characteristics (physicochemical or (micro)biological properties) through conventional analytical chemistry are an immediate first step into working towards this paradigm. Statistical analysis and modeling of this data can be performed to further enhance the use of raw material in process and learn to better classify different lots of the same raw materials.[6] However, the measurement of raw material quality may be time-consuming and prohibitive for in-production monitoring and optimization applications.[7]

This limitation can be overcome by coupled use of advanced spectroscopic methods such as near-infrared (NIRS), UV-visible and Raman spectroscopy with chemometric techniques. This combination constitutes a process analytical technology (PAT) tool. The PAT initiative has been defined by the US Food and Drugs Agency as "(...) a system for designing, analysing and controlling manufacturing through timely measurements (i.e. during processing) of critical quality and performance attributes of raw and in-process materials and processes, with the goal of ensuring final product quality".[8] In this context, it is also common to introduce PAT tools as a fingerprinting and classification of raw materials.[4,9,10]

Pectins are anionic polysaccharides extensively available at the cell walls of fruits. These polymers are predominantly used in the food and beverage industry due to their gelling and thickening properties. Structurally they can be divided into three domains: homogalacturonan (HGA), rhamnogalacturonan-I (RG-I) and rhamnogalacturonan-II. The commercialized

hydrocolloids commonly are constituted by blocks of HGA and RG-I domains. HGA domains are linear sequences of partly methyl esterified α (1→4) linked D-galacturonic acid residues (forming a "smooth" backbone). Whereas in RG-I domains this backbone is interrupted by α (1→2) linked rhamnoses where side-chain groups attach, mostly composed of β (1→4) linked D-galactose or α (1→5) linked L-arabinose.[11,12] The portfolio of pectin applications and its global market is rapidly increasing and with this the need to make a better use of the raw material without compromising end-product quality.[13]

Citrus peels are a rich source of pectin and are the primary raw material for the industrial extraction and purification of this biopolymer.[14] The peels originate as a side-stream of the traditional juice industry, meaning the manufacturers are subjected to scarce, seasonal and poorly standardized supply. This intermittency, together with increasing market demand, has drastic consequences in the feedstock quality variability the pectin production has to withstand.[15,16] Different pectin product applications rely on the different physical and chemical characteristics combinations this polymer can display. The degree of esterification ($\%DE$) and the intrinsic viscosity (IV) are critical quality attributes for the gelation properties of the pectin product. They are related to pectin application in consumer goods, i.e., jam ($5 \leq IV \leq 6$ dL/g, $\%DE \geq 70\%$) and jelly ($IV \geq 6$ dL/g , $58 \leq \%DE \leq 65\%$).[17,18] The values of these properties when the molecule is in its native environment in-peel are higher than at their end-product form, since will progressively suffer degradation effects throughout the extractions process.

These attributes in-peel, together with the concentration of pectin and protopectin, constitute a critical material attributes profile for the raw material. They frame a peel quality profile and are nowadays inferred, in feedback fashion, from data profiles from actual standard extractions. The assessment of this profile, coupled with a model that relates the end product quality (CQA) to the critical process parameters (CPP) and the contribution from the raw materials, allows the sorting of raw material, robust process optimization or process monitoring.[19,20] Therefore there is a need to understand which peel quality assessment strategies are sufficient for the plant operation purpose and ensure that the raw material data is utilized in a valuable and efficient way.

The present work presents a quantitative and qualitative comparison of three different

raw material quality assessment approaches. Two are based on the historical statistical determination of: 1) expert-knowledge defined class, 2) unsupervised clustering classes, and the third is 3) PAT-based with the application of spectroscopic/chemometric methods for the prediction of quality parameters in the raw material from near-infrared spectra. The approaches are performed in the same measurement context, using the same dataset. The paper is organized as follows: first, a description of the materials and methods applied, followed by a presentation and explanation of the different raw material quality assessment approaches, within the citrus peel and pectin extraction case. Then a comparative discussion of the approaches is made, and the last section concludes this study.

Materials and Methods

The following material resources were utilized in light of the opportunity of performing a comparative analysis on a relevant dataset from an industrial partner. The analytical methods are in-house standard procedures while the spectroscopic and chemometric methods are chosen considering their adequacy and widespread presence in the industry.

Citrus peel samples

A total of 85 raw material samples are considered, with Lemon being the most represented fruit (43 samples), followed by Lime (27 samples) and, Orange (15 samples). The citrus peel material in its production form was ground, and no further pre-treatment of the samples was performed. CP Kelco Aps kindly provided all sample material.

Chemical reference

CP Kelco Aps. provided the analytical data for the different physical-chemical variables. A series of lab-scale extractions, both with and without the addition of acid, was performed at standardized conditions ($T = 70$ °C , 3 grams of peel and, for the acidic extractions, 150 mL of 49 mM nitric acid). For each extraction, a total of two samples are taken at $t_1 = 20$ min and $t_2 = 240$ min, respectively. Samples were centrifuged at 10000 g for 10 minutes, and the

supernatant diluted 1:10 with 0.3 M lithium acetate ($pH = 4.6$). The filtrate dilutions were analyzed for pectin concentration (C^{bulk}_{pectin}) and intrinsic viscosity (IV) in a flow injection polymer analysis (FIPA) system. This system comprises a size exclusion chromatography column (1507.8 mm, Thermo Fisher Scientific, MA, USA) with a triple detector array (TDA 305, Vicotek Corp., Houston, USA). A 0.3M lithium acetate ($pH = 4.6$) solution is used as eluent with a flow rate of 1 mL/min at 37 °C. The degree of esterification ($\%DE$) was determined by the 1H NMR method by Winning et al.[21].

Near-infrared spectroscopy

Fourier transformed (FT) near-infrared (NIR) data were collected, on the citrus peel samples, using an ABB Bomem MB3600 FT-NIR technology spectrophotometer (ABB Bomem, Quebec, QC, Canada). The instrument was equipped with a rotating sample module with a quartz window. Spectral data for each sample were collected as the average of 62 single beam spectra at room temperature. The spectra were referenced against a white background spectrum (average of 62 scans). Samples were scanned over a 1000-2632 nm (resolution 16 cm^1) range with a time interval of 20 seconds of measurement. Due to the presence of unwanted and unnecessary physical phenomena capture by the spectra it is often necessary to use preprocessing techniques in spectral data. The techniques provide mainly light scattering and baseline corrections. Spectral derivative techniques also help weighting more importance to otherwise unidentifiable variations within the spectra (e.g. small concentrations of analytes). Mean centering is always used after a sequence of preprocessing techniques.[22]

Additionally, adequate variable selection (wavelength/wavenumber selection) is performed in order to achieve classification and predictive models with better performance. This selection prevents the capture of undesired variance between samples, as well as less redundant information, which will allow for less complex models (fewer latent variables needed to capture the necessary variance). Many different methods exist to assess the best variable selection.[23] One popular technique, which is used in this work, is called interval partial least squares (iPLS). This technique developed by Norgaard et al.[24], builds separate local models on a number of (non-overlapping and equal width) subintervals of the full spectrum region. The regions are then selected based on the prediction performance of these

local (and full-spectrum) models, by means of comparison of the root mean squared error of cross-validation.

Chemometric modeling

The data acquired in the food industry are quite complex and of different nature, often acquired from instrumental measurements comprising thousands of variables (e.g., each wavelength of spectra) for each sample. The complexity of the data has led to the need of diversifying the exploratory methods and employ multivariate data analysis to understand the information in the data. Chemometric methods can be put as a toolset of statistical techniques to analyze datasets with more than one variable (or type of variable). These variables are used simultaneously to perform exploratory (unsupervised learning), regression, or classification (supervised learning) analyses. All the methods considered in this paper have been exhaustively described and published elsewhere (e.g., Massart et al.[25], Sun[26]); therefore, this subsection contains a summarized description of the chemometric methods employed to clarify and provide the reader with theoretical background and appropriate references.

Principal component analysis (PCA)

PCA is a method for dimensionality reduction of large multivariate data sets (i.e., spectral information), useful in many applications in the bioprocessing industries.[27] In essence, PCA is a bilinear decomposition technique that, for a set of observations, relies on an orthogonal transformation of possibly correlated variables into a set of linearly (and mutually) uncorrelated variables called principal components. The transformation is defined such that the first principal component accounts for as much variance in the data as possible. The succeeding components will have the next largest variance, under the constraint that it is orthogonal to the preceding components. It summarizes the observations information in fewer new variables. These new variables, named principal components, are thus composed of linear combinations of the original variables and constitute a set smaller than the original variables set. An original data matrix ($\boldsymbol{X}$), is decomposed into a score matrix ($\boldsymbol{T}$) and a loading

matrix ($\boldsymbol{P}$), with the residuals collected in a matrix ($\boldsymbol{E}$): $\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^T + \boldsymbol{E}$. The loadings define the new coordinates system, the weights that the previous/original variables have on each principal component.[28] The scores are the "amount of" those new artificial variables represented in particular sample, in other words, they are the coordinates of the samples in the principal component space.[29]

Partial least squares regression (PLS-R)

PLS regression has been extensively used for multivariate regression modeling, especially applied to rapid spectroscopic measurements calibration with slow physical-chemical data.[22] In a spectroscopic PLS-R application the purpose is to build a linear model between the desired response variable (y) and the spectrum ($\boldsymbol{x}$), while concurrently maximizing the covariance between them by simultaneously decomposing the predictor and the response matrices iteratively into a reduced set of uncorrelated latent variables (LVs), thereby eliminating redundancy in the datasets.[30] All models have been externally validated, by splitting the dataset into calibration and validation sets using the Kennard-Stone procedure (66% calibration; 34% prediction), keeping the NIR sample replicates together.[31]

Partial least squares discriminant analysis (PLS-DA)

PLS-DA is a supervised classification method, where there is a requirement of prior knowledge, (i.e., the categories of samples). For this model identification, similarly to the PLS-R, a training set of samples for which the categories are known is necessary. This classification method is linear and based on the PLS algorithm, however, modified to perform classification. The main difference is related to the dependent variables (y) as these in PLS-DA are qualitative variables.[32] In PLS-DA, a $\boldsymbol{Y}$ variable matrix is defined as a "dummy variable" matrix and it has as many columns as the number of classes. The information about the class of each sample is provided through a binary code: all entries of each row (corresponding to a sample) are set equal to 0, except for the column corresponding to the category the sample belongs to, whose element is set equal to 1. If there are more than two classes, PLS-DA uses the algorithm PLS2. For each sample a prediction vector, with the size of the number of classes, is retrieved with values close to 0 and 1. The class is determined either

by the maximum value in the y vector or by appropriate threshold setting for each class. Therefore, as in PLS regression, y dependent variables are predicted and thresholds (i.e., for a 2-class problem $y > 0.5$) can be defined to assign the sample to a corresponding class.[33,34] In PLS-DA the focus is not on the prediction error of the model but more the percentage of misclassifications obtained.

Cluster Analysis

Methods in this group of analysis have a common goal of finding groups/classes within a dataset, in which its members share more similarity to each other than with the rest of the observations, not in that group. An extensively used type of cluster analysis is hierarchical cluster analysis (HCA), which is based on distance measurement connectivity between observations. HCA is an unsupervised learning method, which can be agglomerative or divisive. An agglomerative method begins with each sample as its cluster and progresses agglomerating existing clusters into larger ones. Divisive methods start with a single big cluster containing all observations and are continuously separated progress by dividing existing clusters into smaller ones. All these methods require a distance measure between observations, with two popular ones being the Euclidean distance and the Mahalanobis distance. The latter is appropriate to account for multivariate directions. The different algorithms differ in the way the distance between existing clusters (inter-cluster distance) is defined and the decision guide for joining clusters (linkage rule).[28]

Software

All chemometric calculations were performed using Matlab ver. R2015B (Mathworks, Inc.) installed with the PLS Toolbox ver. 8.1. (Wise & Gallagher; Eigenvector Technologies).

Raw Material Quality Assessment Approaches applied to Citrus Peel

There are no specific regulations for the assessment of raw materials, but there are guidelines in the context of good manufacturing practices of active pharmaceutical ingredients. The guidelines push for the evaluation of quality and establishing of acceptance criteria in raw material, correct labeling and documentation of end-product deviations to detect changes resulting from modifications in raw materials.[35] An overview of five different possible approaches for quality assessment of peels is illustrated in Figure 1, of which the last three (highlighted with a dashed line) will be assessed comparatively in a quantitative analysis on their uncertainty. The scheme shows the different approaches regarding the statistical information that can be gathered as well as a qualitative time-effort relative comparison regarding development, as in how much time it is necessary to allocate before being able to use the approach, and in-operation use, as in how time expensive is the approach during production. Figure 1 has elements specific for the pectin case, but transversal for other bio-raw material cases where key parameters of the raw material are identifiable and PAT is applicable. This section will discuss each approach in the context of the articles case study, addressing the necessary developments for each approach.

Rule-of-thumb operation

Historically, the industry has dealt with raw material quality fluctuation in a heuristic manner, relying on the vast experience of their manufacturing teams. Dialogue with raw material suppliers and process engineers has been essential in this way of manufacturing. It requires many years of manufacturing and supplier-relationship for the process engineers to have a "finger on the pulse" and an intangible knowledge of what quality to expect from a certain supplier. This approach gives an "idea" of how to decide the process settings, but it is not fail-proof. Based on how certain categories (i.e., type of fruit, supplier) performed historically, the manufacturer is inclined to select a particular supplier for a particular pectin grade. Figure 1, in the first column, highlights qualitatively the lack of quality variables

determination and the long period required to acquire the experience necessary to operate satisfactorily. Once the expertise exists, the guidelines on how to operate for a specific peel can be readily applied, which means a lower in-operation time effort. In this approach, the raw material quality is still unmeasured and thus uncontrolled noise factor in the whole process.

For instance, the perception that lime peels provide the highest quantity of pectin and the higher end-product IV or that lemon peels typically have a higher end-product %DE, may provide the manufacturer with a false sense of security when purchasing raw material.[16] These citrus peels are purchased depending on the availability and price from various countries and suppliers, with the same suppliers displaying significant fluctuation in manufacture performance for different cultivar years. When using dried citrus peels, the lot variation is reduced compared to that of a fresh fruit lot. However, even though situations may arise where there is apparent robustness (especially if the raw material fluctuation is not severe) this approach provides no statistical information about the raw material. The producer will not be able to cope accordingly in the face of eventualities (e.g., supplier bankruptcy) that require a sudden change of the main supplier, or even if the raw material quality of a trusted partner changes due to changes in their protocols.

Wet-Lab Analysis

The following approach in Figure 1 requires the determination of critical (raw) material attributes measurable via experimental analysis. The development of an analytical method starts with the definition of the desired characteristics to assess in a substance. Once identified the relevant critical material attributes, the analytical procedure depends on the choice of sample preparation, analytical instrumentation and methods that are appropriate for the nature of the sample and the intended goal of the analysis.[36] After development, it is necessary to evaluate the method by setting appropriate acceptance criteria in validation experiments for typical parameters, such as specificity, linearity, accuracy, precision, range, detection limit, quantitation limit, and robustness.[37] Although bio-analytics are fully regulated in the pharmaceutical industry[38,39], the food and bio-based fields lack detailed guidelines for analytical method validation.

The process of identifying the desired attributes, development of the analytical method and validation can be long and require a significant amount of experimentation to achieve robustness. However, one can expect to achieve an acceptable method in less time than it takes to identify consistently (if possible) the quality distribution in the supplier market by qualitative feedback from production (approach in the *Rule-of-thumb operation* section). While in operation, these methods can be extremely time and resource consuming (e.g., personnel, reagents), with the results, typically, only accessible after hours (even days), preventing the use of the information obtained for control of the current process. The estimate resulting from a bio-analytical method is expected to have a relative standard deviation of less than 15%.[38]

Determining raw material attributes that are parameters (or initial states) of a model is highly useful as it enables their direct use. The outputs of the analysis would be inputs in the model, creating a flexible tool for simulation of key performance indicators for a given raw material and process settings. Different models for pectin solid-liquid extraction have been developed[40–43], but lacked the integration of critical material attributes in the process dynamics. Andersen et al.[44] have developed a model which describes the relevant process key performance indicators, pectin concentration in the bulk media (C^{bulk}_{pectin}) effect of temperature and pH and incorporates parameters which are both related to the concentration, C^{0}_{pectin} and $C^{0}_{protopectin}$, and initial quality, IV_0 and $\%DE_0$ of pectin material in the peels. These parameters, $\boldsymbol{u_{peel}} = [C^{0}_{pectin}, C^{0}_{protopectin}, IV_0, \%DE_0]$, constitute a vector of critical material attributes which will vary from peel to peel. The experimental procedure and analytical methods described in the *Chemical reference* section were developed as a means to measure these parameters directly:

- $C^{0}_{pectin} = C^{bulk}_{pectin}(t_2, \text{water extraction})$, where t_2 is the second sampling time and, assuming only readily available pectin is extracted with water.
- $\%Yield = \frac{C^{bulk}_{pectin}(t_2, \text{acid extraction}) \times V_{extraction}}{m_{peel}} \times 100$, which is not explicitly a $\boldsymbol{u_{peel}}$ parameter, but it is used as a variable in the results for the ease of dimensionless comparison between samples.

indirectly:

- IV_0 and $\%DE_0$ by assuming a first-order reaction (i.e., $ln[IV] = ln[IV_0] - k \times t$ and calculating the y-intercept with the IV and $\%DE$ data at t_1 and t_2.

moreover, by reconstruction:

- $C^0_{protopectin} = C^{bulk}_{pectin}(t_2, \text{acid extraction}) - C^0_{pectin}$, reconstructed from the measured variables above and thus not used in the results below.

Performing these assays will provide us the uncertainties associated with the measurements for each peel. Lab-scale extraction methods are favored in detriment of the pilot extractions. It is (manifold) less resource and time intensive, and the extraction conditions can be carefully standardized at lab scale (e.g., thermal bath, perfect mixing, and better sample handling). However, these experiments are still time and resource consuming methods. This is undesirable for a standard operating procedure, aimed at systematic use in production. There is a need to streamline the assessment of peel quality, but maintaining the statistical information essential for further applications.

Historical dataset statistics

Once the critical material attributes are determined, the method is established, and data has been continuously collected it is possible to characterize the raw materials statistically. If the information is gathered on a representative population, including prospective raw materials not currently in production, the approach offers a robust variation assessment of the tested materials. The crucial point in this approach is the need to guarantee that the samples tested do cover the expected variation in the production environment. When established, information on the critical material attributes of incoming raw material can be inferred from the up-to-date dataset statistics, and routine sample analysis can be made to update the dataset. Building this dataset, as the third column in Figure 1 illustrates, solves the in-operation time pitfall of the previous approach. However, the process to ensure representativity can be lengthy and resource consuming there is no guarantee that the extremes are investigated, and future samples are out of the analyzed limits.[5]

Table 1 offers the statistical summary of the measured variables in-peel. Due to the happenstance nature of the dataset, not all samples have a complete critical material attributes

vector ($\boldsymbol{u_{peel}}$). Each variable in Table 1 has indicated the number of samples, n, which got measured for that particular variable. The summarization of the data is assessed using empirical distributions: a measure of the central tendency (mean, median, mode); a measure of spread (range, quartiles, standard deviation); a measure of asymmetry (skewness) and peakedness (kurtosis) of the data distribution. Skewness measures the lack of symmetry in distribution. A variable is symmetric if it looks the same to the left and right of the center point. Kurtosis measures how tailed a distribution is relative to a normal distribution. A normal distribution is symmetric with well-behaved tails. For this type of distribution, the skewness is close to 0 and kurtosis has the value of 3.[45] Distributions with higher kurtosis will have heavier tails (possible outliers). The variables in Table 1 display similar behavior to a normal distribution. However, operating based on the complete dataset statistics could be flawed and does not provide the flexibility to account for significant raw quality variations which may occur. The uncertainty associated with the raw material might be too large to consider a single statistical distribution for valuable implementations.

In the citrus peels example, this is known a priori because we know beforehand that different fruits are typically used to produce pectins with different specifications.[14] A mean of 7.78 and standard deviation of 1.14 for IV_0 places the manufacturer's initial guess for a given peel within the operational range of the two different product-specifications previously mentioned: jam ($5 \leq IV \leq 6$ dL/g) and jelly ($IV \geq 6$ dL/g). However, many peels in the dataset have a measured $IV_0 < 6$ dL/g and the negative skewness (-0.28) indicates that the distribution for IV_0 has higher incidences for values $IV_0 < 7.76$ dL/g. This would be problematic if we adopted the global mean of the dataset value and attempted to produce jelly-type pectin.

Class dataset statistics

The shortcomings in the previous section motivate the search for classification that yields a narrower window of uncertainty regarding the critical material attributes. Moreover, they can provide the manufacturer with specific operational guidelines based on the class of the incoming raw material, connecting different optimal operating setting to different raw materials. These classes can be defined with either qualified expert-knowledge (discriminant)

classification or through unsupervised learning and clustering algorithms.

Expert-knowledge classification

Expert-knowledge classification relies on prior information, which the manufacturer uses to sort different samples, for example, attributing classes to raw materials based on their supplier, country of origin and in this case the type of fruit. Figure 2 shows what this classification yields in statistical terms using boxplot visualization. This graphical method provides additional information to the summary statistics. It can identify outliers, changes in the data distribution across different groups and variables and even highlight relationships between variables. As it can be seen in Figure 2, there are a few outliers in the selected groups and variables, the most noteworthy being the orange sample which has outlier values for IV_0 and $\%DE_0$. This is a case where it is possible that an error while compiling the dataset occurred in labeling the sample, as it is evident that this sample has a very distinct value in variables where its group (orange) is well differentiated from the rest. The sample could have been a lime or lemon, with a higher probability as the former since it most closely fits the lime interquartile range (the box in Figure 2 that represents 50% middle) in all variables.

The boxplots highlight the dissimilarities between fruits, with some variables having more discernible differences between fruits than others do. For instance, the IV_0 and $\%DE_0$ boxplots clearly isolate orange and the rest (see Figure 2). Lemon and lime, albeit overlapping in their data range, still have different central differences. The lower IV_0 from orange supports the study by Kaya et al.[46] that claims oranges contain longer or more numerous side stretches which provides a flexible conformation, leading to decreased intrinsic viscosity values compared to lime and lemon. For these quality variables, the fruits reveal an almost normal distribution (mean and median close together, and the whiskers have similar lengths). This is not the case for initial pectin content in the peels.

For the C^0_{pectin} variable, data distribution by fruits overlaps significantly with no highly discernible central distribution difference between fruits. However, Figure 3 shows visible clustering by the supplier. This indicates that the peel pre-treatment of each supplier is a defining factor, rather than the fruit group itself. This is in agreement with what is

stated in the literature, that the pectin solubility is promoted by a combination of the de-esterification of the polygalacturonic acid backbone (pectin methylesterases) facilitating the depolymerization of pectins (polygalacturonase) and the cleavage of linkages between side chains of pectin and hemicelluloses. This conversion of protopectin to soluble pectin is dependent on the pre-treatment (blanching, washing, drying, etc.) applied by the supplier.[18]

The different stages of fruit maturity, when the peels are collected and sold by the supplier, also contribute to the differentiation between pectin in the different peels.[11] Adopting a classification based on the fruit type is a lesser than an optimal solution since for the $\boldsymbol{u_{peel}}$ variables there is excessive overlap between the fruit types. This motivates the examination for classes that are "blind" to the fruit type.

Unsupervised learning and cluster analysis

Unsupervised clustering is performed with little or no information about class structure before the classification; the classes form based on the distance of the $\boldsymbol{u_{peel}}$ vector between samples. A PCA model, based on the measured critical material variables, with two components (84% cumulative variance), is built and analyzed such that the loadings and scores can be visualized and can be interpreted together in a biplot.[47] The PC1 vs. PC2 bi-plot is seen in Figure 4. The first principal component is able to make a separation between the three fruits. This confirms what was previously assessed in the univariate analysis, that the fruits were distinguishable to a certain degree with the $\boldsymbol{u_{peel}}$ variables. The fruits are mainly separated by the IV_0 and $\%DE_0$ variables. However, there are lemon and lime samples that the PCA model cannot tell apart. This is in accordance with the previously assessed information on the overlaps between lime and lemon. Other information extracted from Figure 4 is that the IV_0 and $\%DE_0$ variables are near each other and far from the origin, meaning they are correlated (with respect to the variation explained by the components). Cluster Analysis of these scores will allow for establishing classes that respect the closeness of data in a multivariate sense, irrespectively of their fruit type.

In Figure 5A, the classes originated from using Wards agglomerative algorithm method[48], using Mahalanobis distance, on the PC1 and PC2 scores from the previous PCA model. This approach generates more homogeneous classes, which are not overlapping in the PC1 vs PC2

plane, contrary to the fruit classification. Class 1 has both lime and lemon samples as their constituents. Class 2 is comprised of only lemons, which differentiate themselves from the rest for their high IV_0 and $\%DE_0$ values. Finally, class 3 coincides entirely with the orange samples, which are known from the univariate analysis to be distinguishable from the other fruits, particularly in terms of IV_0 and $\%DE_0$ values. It is a spurious classification attempt since the experimenter can identify if a sample belongs in class 3 merely by identifying the sample as an orange. This encourages the clustering in Figure 5B within the lemon and lime samples, in an identical mode to the previous clustering but without the undesired variance that is captured by adding the orange samples. The classes in the new analysis separate once more the samples which have high IV_0 and $\%DE_0$ values (class 2) from the others. This yields class 1, with mixed samples of lemon and lime, while class 2 is exclusively composed of lemons. To determine the class of a sample the determination of its $\boldsymbol{u_{peel}}$ vector (or partly) is required, which can still be cumbersome for the operational decision-making and optimization in a timely manner. Any monitoring or process optimization strategy will opt for an approach that provides the best estimates (with the least uncertainty) in a rapid manner. This encourages coupling the critical material attributes, $\boldsymbol{u_{peel}}$, with a spectroscopic method.

Spectroscopic coupling

Spectroscopic techniques provide the manufacturer with a fast and, in most circumstances, non-invasive and non-destructive tool. The use of spectroscopic techniques can be fruitful in reducing drastically the in-operation time when compared to wet-lab analysis, and providing the manufacturer with a better estimate of the critical material attribute than the dataset statistics. This is illustrated in the last column in Figure 1. These tools can also reduce the time-effort put into developing the quality assessment approach if used for the initial screening of the different batches of raw material. The analytical tests can then be made in a reduced set of the original selection, carefully selecting the materials that cover the largest variation ensuring a representative dataset. In the following examples, the methods are used in the context of the full-dataset employed in the previous approaches. Within the possible techniques, NIRS is a notably reproducible and robust spectroscopic method that

has proved its rapid non-invasive use across the food and agrochemical industries. Previous studies register NIRS capabilities for detecting pectins and pectin quality parameters.[49,50]

Principal component analysis is performed to investigate the samples separation based on their full NIR spectra information. The score plot for the two first components is shown in Figure 6, with the spectra being pre-processed with the common standard normal variate (SNV) and mean centering techniques. Pre-processing attempts to remove physical variability (scatter correction), so that the samples can span in the principal components space due to variations in the chemical matrix. The first component explains 83.7% of the variation in the pre-processed NIR samples. By analyzing the PCA score plot in Figure 6A, it is possible to observe a discernible gap between samples in PC1 (highlighted with a red box). The reason for this can be assessed by evaluating the spectra, in Figure 6B, together with the loadings plot for PC1, in Figure 6C. A wavenumber range which has high impact in the separation across PC1 is 8400-10000 cm^1, in the third overtone NIR region. Not only it is the region with the weaker intensity it can also be seen that it manifests specular reflectance effects with no sharp structure. The first component is largely composed of effects resulting from differences in sample pre-treatment (grinding, storage, etc.) rather than differences in the pectin molecules and its availability in the peel. This large variation captured only adds noise to the purpose of using the spectra to infer $\boldsymbol{u_{peel}}$ information from a samples spectrum. This is an indication that this region should not be used for further classifications or predictions of the physicochemical variables of pectin in the peel matrix. In succession, the second component has dominant loadings in a band that corresponds to RCO2R ester-groups (maximum at 5161 cm^1). This is a good indication that the samples are separated by their degree of esterification in this direction. It is important to note that the replicates are also close to each other, indicating a good degree of robustness. Another interesting feature is that the physical effect is not exclusive for a single fruit (1 lemon sample) and only the suppliers "1", "3", "10" and "13" are affected, but rather than being a characteristic spectroscopic fingerprint from the supplier, it is possible that it is merely the effect of all these samples having been pre-treated similarly. One evidence for this possibility is that it is possible to find other samples of supplier "13" outside the highlighted box in Figure 6A. In these components, samples from the same suppliers do not necessarily occupy the same

regions of the PC space. For example, this is also visible for the supplier "5".

Classification aiding tool

PAT based classification can overcome the analytical effort pitfall, through spectroscopic class assignment or spectroscopic prediction of classes assigned based on reference analysis measurements. Based on the unsupervised classes obtained from the cluster analysis in the *Unsupervised learning and cluster analysis* section, a PLS-DA model attempts to predict these classes (obtained with the reference data proximity) only by using the NIR spectra of the samples. By predicting the class to which a sample belongs to, the statistics (mean, standard deviation) from that class can then be used as the $\boldsymbol{u_{peel}}$ vector for that sample, the same way the statistics for the supervised classes in "Fruits" would be applied. The spectra are preprocessed with the same techniques applied in the analysis above, standard normal variate (SNV) and mean centering. Cross-validation is applied by leaving out the NIR replicates of the same peel together, thus avoiding over-fitting of the data. A confusion matrix is used to assess the PLS-DA classification success in classifying the samples, illustrating the correctly classified samples, type-I errors (false-positives), and type-II errors (false-negatives) samples for each category in the matrices presented in Table 2. It can be observed that 16 (Class 1) and 18 (Class 2) samples were misclassified by the model attempting to classify all samples. It can correctly identify class 3 samples (oranges), but there are errors on samples close to the intersection between the two classes containing both lemons, and limes. This indicates that the model is not perfect (fail proof efficient) in the classification process. However, a practical implementation for our case, making use of a priori information of the fruit type, would be to classify the lemons samples between the two classes in Figure 5B. This reduces the task complexity of the model dramatically and yields an almost perfect cross-validated result, seen in Table 2. The two misclassifications are replicates of the same peel, with a third replicate having been correctly classified. Furthermore, by performing a variable selection with an iPLS algorithm the cross-validated result is improved for both classification models, with the latter achieving a perfect cross-validated result.

Critical material attributes prediction models

If the raw material key parameters are liable to be calibrated with a spectroscopic tool, an at-line PAT application can provide that timely information. The calibration makes use of the acquired dataset in the attempt, of quantitatively predicting the $\boldsymbol{u_{peel}}$ variables. The same calibration and external validation sets were used for all models. The results from the $\%DE_0$ model, illustrated in Figure 7, showcase the adequacy of the NIRS technique for the characterization of the specified material attributes. This model was built in an iterative fashion, trying different combinations of spectra pre-processing and variable selection. The outcomes from the different models are compared based on their performance to predict the same external validation set. This process follows these steps:

1. Choosing the pre-processing of spectra (i.e., SNV + 1^{st} derivative + mean centering) and lab-reference $\%DE_0$ samples (i.e., autoscaling)
2. Variable selection (i.e., no variable selection: full spectra)
3. Number of latent variables selection. Register the model performance.
4. Re-do model with subtle changes to variable selection (i.e., use iPLS algorithm). Register the model performance.
5. Go through sequence 1-4 with changes to 1) (i.e., SNV + 2^{nd} derivative + Mean Centering), register the model performance.
6. Comparison of root mean square error of prediction (RMSEP) of the external validation set

From the different modeling iterations, the $\%DE_0$ model that yielded the best results was obtained for the regression with a pre-processing of the spectral data with SNV + 1st derivative + mean centering. The R^2 is the squared correlation coefficient providing the explained y-variance and $bias = (y_{pred} - y_{ref})/n$, with n as the number of samples. There is a high correlation in the model and the normalized-RMSEP ($nRMSEP = \frac{RMSEP}{(X_{\mathrm{obs,max}} - X_{\mathrm{obs,min}})} \times 100$) is close to 10%. There is, however, a relatively high bias, which results from a poorer predictive capability of the model for orange peels. It should be pointed out that at this point,

the distribution of the samples (see *Citrus peel samples* section) considered for this model building is somewhat skewed and might not allow for a generic model that predicts critical material attributes for all peels adequately. Further efforts will go into validating and consolidating the results. Nonetheless, the study showed in principle that the technique is capable of prediction.

Quantitative statistical comparison

A quantitative comparison of the three approaches discussed previously i.e., fruit classes (*Expert-knowledge classification* section), cluster classes excluding the orange samples (*Unsupervised learning and cluster analysis* section) and the partial least squares regressions (*Critical material attributes prediction models* section) is shown in Table 3 and Figure 8. The comparison is made by a measure of central tendency (i.e., the mean μ) and a measure of spread (i.e., the standard deviation σ) within the same dataset context. When taking the mean of a variable as the prediction for a certain class, the uncertainty on such assumption can be defined by the standard deviation of such variable in that class. For example, if the manufacturer operates based on fruit discrimination and a lemon peel arrives in production, they can assume the values in the vector $\hat{\boldsymbol{u}}_{\textbf{incoming lemon peel}} = [8.4; 75.1; 1.72; 23.69]$ as the critical material attributes profile for that raw material. It is then possible to optimize the process stochastically according with the desired product specifications, knowing that the uncertainty associated to such assumption can be defined by the standard deviations of the $\hat{\boldsymbol{u}}_{\textbf{incoming lemon peel}}$ variables: $\boldsymbol{\sigma}_{\hat{\boldsymbol{u}}_{\textbf{incoming lemon peel}}} = [0.98; 1.98; 0.51; 1.57]$.[19] It can be assessed in Table 3 that there is a statistical uncertainty improvement from fruit classes (qualified expert-knowledge) to cluster classes (unsupervised). Across all $\boldsymbol{u_{peel}}$ variables, the cluster classes have a smaller maximum standard deviation, when compared to fruit classes. The uncertainties associated with the model predictions are well in the range (i.e., IV_0 and $\%Yield$), or smaller (i.e., $\%DE_0$ and C^0_{pectin}) than the minimum standard deviations obtained in the group statistics. In Figure 8, the mean value of the relative standard deviations (also known as coefficients of variation) of the composing classes of a given approach is calculated for each $\boldsymbol{u_{peel}}$ variable.

Discussion

The quantitative comparison showed how both statistics of historical analytical data and the performance of predictive spectroscopic models are useful to assess raw material quality with an appropriate uncertainty. The manufacturer can opt to rely on the class statistics or implement the PAT tools, depending not only on performance but also on practicality and economic feasibility. For the fruit-based classification, experimentation can be skipped once a representative dataset is available. Samples are classified given an intrinsic characteristic which is known *a priori*. Classification of samples based on their original fruit provides a decent distinction and requires no classification model. However, it shows to be insufficient for specific critical material attributes, i.e., C^0_{pectin} (see Figure 3), which show an overlap of different fruits.

Alternatively, classification of samples based on cluster analysis joins samples in groups based on their multivariate proximity regarding the identified critical material attributes. This yields classes that are more homogeneous concerning uncertainty (see Figure 5 and Table 3). This approach makes the most of all available wet-lab features of the samples and is not influenced by biased classifications based on heuristics. This results in classes which include a mix of fruits. However, this type of classification requires additional information. Performing the full lab experimentation would allow for this classification, but is also beneath the purpose of bypassing the wet-lab analysis. A way to partially circumvent this is to attempt classification through partial experimentation and posterior determination of the lacking variables through correlation or missing-data algorithms.

The *Spectroscopic coupling* section explores the use of NIRS as a tool for rapid identification of these classes. In a previous study by Engelsen et al. [50], NIRS showed discriminant capability and was able to successfully distinguish citrus peels samples originating from different countries and specific producer fingerprints. A particularly cumbersome difficulty is how the raw material sample pre-treatment affects the multivariate applications. This is observed in Figure 6, where a great part of the variance between samples captured is non-dependent on the chemical matrix of the sample. This is a common case in biological samples where scattering properties are complex. [22] Caution is necessary for standardizing the pre-

treatment of samples. Different sample grinding could also induce differences in the chemical matrix, specifically in the water content. Moisture loss occurs during grinding, mainly due to air throughput exposure in the particle which has a bigger superficial area. The grinding also exposes the particles to overheating. In this study, the approach with PLS-DA modeling suffers from a few misclassification errors between class 1 and 2 (Table 2). However, a faultless classification is possible when combining heuristic information (knowledge of a samples fruit) and the PLS-DA model on lemon samples. This application has a smaller classification task and is still very relevant in operation as it allows the users to identify via NIR if a lemon belongs to class 1 or class 2 (and expect higher IV_0 and $\%DE_0$ values). With the creation of this class distinction, the user avoids the underestimation, by assuming a fruit class, of IV_0 and $\%DE_0$ values for good performing lemon samples. It provides not only better precision, as seen in Table 3 and Figure 8, but also better accuracy. An alternative approach would involve building classes based on proximity (clustering) of the near-infrared samples PCA scores. This would enable better performance of a classification model based on the NIRS. However, the samples would still need to be analyzed in wet-lab, as to give the manufacturer the statistical information of variables belonging to each NIRS class. This would possibly yield classes less homogeneous concerning wet-lab attributes uncertainty but could form classes which comprise more latent information, than the ones performed in this study. Additionally, different classification model algorithms such as the more traditionally used soft independent modeling of class analogy (SIMCA)[34,51]; or the recent trend of ensemble methods (e.g., Random Forest) have to be explored, as they might yield better performances.[52–54]

It is shown that NIRS can also characterize the citrus peel raw materials with PLS-R predictive models of three variables of $\boldsymbol{u_{peel}}$ (IV_0, $\%DE_0$ and C^0_{pectin}) and providing a reasonable estimate for $\%Yield$ (Table 3). For an unbiased estimator model, as in the PLS-R example, the root mean square error of prediction (RMSEP) is equivalent to the standard deviation. The model predictions for$\%Yield$ exhibit a poorer correlation, understandably so since it is a material attribute that is less explicitly related to direct correlation in the spectra, and depends more on the full extent of a test extraction. However, a RMSEP=1.57% is still within the uncertainty range of the other methods and this approach allows having

a quick estimate, not reliant on wet-lab analysis of the sample in question. Additionally, these predictions provide a more accurate central tendency: a sample which would be in the extreme of a group statistic, thus with the group mean being a bad estimation of its true value, is individually predicted by the PLS-R model.

It should be stressed that the comparison of approaches in this study is made under the premise of negligible noise in the reference values compared with both the class uncertainty distribution and the prediction uncertainty. However, this is not always the case, especially when dealing with biologically derived raw material where the biotic noise may severely influence the accuracy of the reference method. This is an additional point in favor of using NIRS, as it has been shown that predictions for a group of samples can be closer to their true values than the set of lab analysis for this same group of samples.[55] In fact, the RMSEP calculated in this study (Table 3) are in reality the apparent RMSEP, dependent on both the errors in the lab values and the inherent model errors and can be a pessimistic expectation of prediction uncertainty. An effective correction for the reference error component leads to $RMSEP_{corrected} = \sqrt{\left[RMSEP^2_{apparent} - \hat{\sigma}^2_{error}\right]}$ where $\hat{\sigma}^2_{error}$ is the estimate of variance of the reference method.[56] This error estimate can be computed with the full analysis of the repeatability and reproducibility (e.g., Gage R&R) of the reference method.[57]

When applicable, PAT tools should be favored as the in-process benefits are manifold. The use of NIRS raw material identification has the significant advantage of enabling at-line analysis directly at the reception in the warehouse or the feed-inlet of the tanks. Another aspect to be considered is that raw material may suffer in-lot variation and wet-lab analytical methods are performed on a minute amount. This can lead to production quality drifts if the lot in question is used continuously and the estimates on the critical material attributes are not adapted to this in-lot variations. A spectroscopic approach would be a low-cost and straightforward strategy to screen this variation continuously. Overall, PAT tools offer significant improvements in speed of analysis and resource wasting, enabling faster decision-making. The use of PAT does not motivate a complete elimination of the wet-lab analysis set-up. The calibration models need to be continuously updated, and the manufacturer needs to ensure new suppliers are within the previously established design space for raw materials, but it alleviates the quality control laboratory from the production optimization support

task and allows for fewer tests.

Conclusions

The work explored different approaches to determining raw material critical attributes in a bio-based context. The path from a heuristic-based to a PAT-based operation is presented, highlighting the key differences regarding time-expenditure (in development and in-operation) together with the information each approach provides the manufacturer. Characterization of dried citrus peels through multivariate data analysis was performed and illustrated the successive developmental nature of the different approaches. An investigation on how the quantitative performance of chemometric near-infrared spectroscopy prediction models compare with dataset class statistics (based on expert-knowledge or clustering algorithm classes) was assessed. The non-invasive spectroscopic method has been proven to have the potential to characterize pectin extraction raw material with minimal sample preparation. The study shows the potential benefits of opting for PAT-based approaches on the early stage of bioprocessing plants when compared with industry standards of "rule-of-thumb" operation.

Acknowledgments

The project received financial support from Innovation Fund Denmark through the BIOPRO2 strategic research centre (Grant number 4105-00020B).

Literature Cited

1. Rathore, A.S.. QbD/PAT for bioprocessing: Moving from theory to implementation. *Current Opinion in Chemical Engineering* 2014; 6:1–8.

2. Lanan, M.. QbD for Raw Materials. In: Rathore, A.S., Mhatre, R., eds. *Quality by Design for Biopharmaceuticals*; chap. 11. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2008:193–209.

3. Mevik, B.H., Færgestad, E.M., Ellekjær, M.R., Næs, T.. Using raw material measurements in robust process optimization. *Chemometrics and Intelligent Laboratory Systems* 2001; 55(1):133–145.

4. Rathore, A.S., Mittal, S., Pathak, M., Arora, A.. Guidance for performing multivariate data analysis of bioprocessing data: Pitfalls and recommendations. *Biotechnology Progress* 2014; 30(4):967–973.

5. Skibsted, E.. Process Analytical Technology Applied to Raw Materials. In: Undey, C., Low, D., Menezes, J.C., Koch, M., eds. *PAT Applied in Biopharmaceutical Process Development And Manufacturing An Enabling Tool for Quality-by-Design*; chap. 7. CRC Press-Taylor & Francis Group; 2011:127–141.

6. Berget, I., Næs, T.. Optimal Sorting of Raw Materials, Based on the Predicted End-Product Quality. *Quality Engineering* 2002; 14(3):459–478.

7. Jørgensen, K., Næs, T.. A design and analysis strategy for situations with uncontrolled raw material variation. *Journal of Chemometrics* 2004; 18(2):45–52.

8. Food and Drug Administration, . PAT Guidance for Industry A Framework for Innovative Pharmaceutical Development, Manufacturing, and Quality Assurance. Tech. Rep.; 2004.

9. Jose, G.E., Folque, F., Menezes, J.C., Werz, S., Strauss, U., Hakemeyer, C.. Predicting mab product yields from cultivation media components, using near-infrared and 2D-fluorescence spectroscopies. *Biotechnology Progress* 2011; 27(5):1339–1346.

10. Kirdar, A.O., Chen, G., Weidner, J., Rathore, A.S.. Application of near-infrared (NIR) spectroscopy for screening of raw materials used in the cell culture medium for the production of a recombinant therapeutic protein. *Biotechnology Progress* 2009; 26(2):527–531.

11. Sriamornsak, P.. Chemistry of Pectin and Its Pharmaceutical Uses : A Review. *Silpakorn University International Journal* 2003; 3(1-2):206–228.

12. Willats, W.G.T., McCartney, L., Mackie, W., Knox, J.P.. Pectin: cell biology and prospects for functional analysis. *Plant Molecular Biology* 2001; 47:9–27.

13. Ciriminna, R., Fidalgo, A., Delisi, R., Ilharco, I.M., Pagliaro, M.. Pectin Production and Global Market. *Agro Food Industry Hi Tech* 2016; 27(5):17–20.

14. May, C.D.. Pectins. In: Imeson, A.P., ed. *Thickening and Gelling Agents for Food*. Chapman & Hall; second ed.; 1997:230–261.

15. Ciriminna, R., Chavarría-Hernández, N., Inés Rodríguez Hernández, A., Pagliaro, M.. Pectin: A new perspective from the biorefinery standpoint. *Biofuels, Bioproducts and Biorefining* 2015; 9(4):368–377.

16. May, C.D.. Industrial Pectins: Sources, Production and Applications. *Carbohydrate Polymers* 1990; 12:79–99.

17. CP Kelco, . GENU pectin Book. Tech. Rep.; 2010.

18. Lopez da Silva, J.A., Rao, M.A.. Pectins: Structure, Functionality, and Uses. In: Stephen, A., Phillips, G., Williams, P., eds. *Food Polysaccharides and Their Applications*; chap. 11. CRC Press-Taylor & Francis Group; second ed.; 2006:353–411.

19. Caroço, R.F., Kim, B., Santacoloma, P.A., Lee, J.H., Abildskov, J., Huusom, J.K.. Analysis and model-based optimization of a pectin extraction process. *Journal of Food Engineering* 2019; 244:159–169.

20. Måge, I., Næs, T.. Optimising production cost and end-product quality when raw material quality is varying. *Journal of Chemometrics* 2007; 21:440–450.

21. Winning, H., Viereck, N., Nørgaard, L., Larsen, J., Engelsen, S.B.. Quantification of the degree of blockiness in pectins using H NMR spectroscopy and chemometrics. *Food Hydrocolloids* 2007; 21:256–266.

22. Rinnan, Å., van den Berg, F., Engelsen, S.B.. Review of the most common pre-processing techniques for near-infrared spectra. *TrAC Trends in Analytical Chemistry* 2009; 28(10):1201–1222.

23. Xiaobo, Z., Jiewen, Z., Povey, M.J.W., Holmes, M., Hanpin, M.. Variables selection methods in near-infrared spectroscopy. *Analytica Chimica Acta* 2010; 667(1-2):14–32.

24. Norgaard, L., Saudland, A., Wagner, J., Nielsen, J.P., Munck, L., Engelsen, S.B.. Interval Partial Least-Squares Regression (iPLS): A Comparative Chemometric Study with an Example from Near-Infrared Spectroscopy. *Applied Spectroscopy* 2000; 54(3):413–419.

25. Massart, D., Vandeginste, B., Deming, S., Michotte, Y., Kaufman, L.. Chemometrics: a textbook. Elsevier Science B.V.; 1988.

26. Sun, D.W.. Infrared spectroscopy for food quality analysis and control. Academic Press/Elsevier; 2009.

27. Skov, T., Honoré, A.H., Jensen, H.M., Næs, T., Engelsen, S.B.. Chemometrics in foodomics: Handling data structures from multiple analytical platforms. *TrAC Trends in Analytical Chemistry* 2014; 60:71–79.

28. Li Vigni, M., Durante, C., Cocchi, M.. Exploratory Data Analysis. In: Marini, F., ed. *Data Handling in Science and Technology*; vol. 28; chap. 3. Elsevier B.V.; 1 ed.; 2013: 55–126.

29. Bro, R., Smilde, A.K.. Principal component analysis. *Analytical Methods* 2014; 6(9):2812–2831.

30. Geladi, P., Kowalski, B.R.. Partial least-squares regression: a tutorial. *Analytica Chimica Acta* 1986; 185:1–17.

31. Kennard, R.W., Stone, L.A.. Computer Aided Design of Experiments. *Technometrics* 1969; 11(1):137–148.

32. Brereton, R.G., Lloyd, G.R.. Partial least squares discriminant analysis: taking the magic away. *Journal of Chemometrics* 2014; 28(4):213–225.

33. Ballabio, D., Todeschini, R.. Multivariate Classification for Qualitative Analysis. In: Sun, D.W., ed. *Infrared Spectroscopy for Food Quality Analysis and Control*; chap. 4. New York, NY: Academic Press; 2009:83–104.

34. Bevilacqua, M., Bucci, R., Magri, A.D., Magri, A.L., Nescatelli, R., Marini, F.. Classification and Class-modeling. In: Marini, F., ed. *Data Handling in Science and Technology*; vol. 28; chap. 5. Elsevier B.V.; 2013:171–233.

35. ICH Expert Working Group, . Q7 - Good Manufacturing Practice Guide for Active Pharmaceutical Ingredients. Tech. Rep. November; 2000.

36. Chauchan, A., Mittu, B., Chauchan, P.. Analytical Method Development and Validation: A Concise Review. *Journal of Analytical & Bioanalytical Techniques* 2015; 6(1):1–5.

37. Green, J.M.. A Practical Guide to Analytical Method Validation. *Analytical Chemistry* 1996; 68:305A–309A.

38. ICH Expert Working Group, . Q2 (R1) - Validation of Analytical Procedures : Text and Methodology. Tech. Rep.; 2005.

39. Tiwari, G., Tiwari, R.. Bioanalytical method validation: An updated review. *Pharmaceutical methods* 2010; 1(1):25–38.

40. Cho, Y.J., Hwang, J.K.. Modeling the yield and intrinsic viscosity of pectin in acidic solubilization of apple pomace. *Journal of Food Engineering* 2000; 44(2):85–89.

41. Durán, R., Villa, A.L., Ribeiro, R., Rabi, J.A.. Pectin Extraction from Mango Peels in Batch Reactor: Dynamic One-Dimensional Modeling and Lattice Boltzmann Simulation. *Chemical Product and Process Modeling* 2015; 10(3):203–210.

42. Minkov, S., Minchev, A., Paev, K.. Modeling of the hydrolysis and extraction of apple pectin. *Journal of Food Engineering* 1996; 29(1):107–113.

43. Pagan, J., & Ibarz, A.. Extraction and rheological properties of pectin from cocoa husks. *Journal of Food Engineering* 1999; 39:193–201.

44. Andersen, N., Cognet, T., Santacoloma, P., Larsen, J., Armagan, I., Larsen, F., Gernaey, K., Abildskov, J., Huusom, J.. Dynamic modeling of pectin extraction describing yield and functional characteristics. *Journal of Food Engineering* 2017; 192:61–71.

45. Tukey, J.W.. Exploratory data analysis. Addison-Wesley Pub. Co; 1977.

46. Kaya, M., Sousa, A.G., Crepeau, M.J., Sorensen, S.O., Ralet, M.C.. Characterization of citrus pectin samples extracted under different conditions: influence of acid type and pH of extraction. *Annals of Botany* 2014; 114(6):1319–1326.

47. Gower, J.. A general theory of biplots. In: Krzanowski, W., ed. *Recent Advances in Descriptive Multivariate Statistics*. Oxford University Press; 1995:283–303.

48. Murtagh, F., Legendre, P.. Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? *Journal of Classification* 2014; 31:274–295.

49. Baum, A., Dominiak, M., Vidal-Melgosa, S., Willats, W.G.T., Søndergaard, K.M., Hansen, P.W., Meyer, A.S., Mikkelsen, J.D.. Prediction of Pectin Yield and Quality by FTIR and Carbohydrate Microarray Analysis. *Food and Bioprocess Technology* 2017; 10(1):143–154.

50. Engelsen, S.B., Mikkelsen, E., Munck, L.. New approaches to rapid spectroscopic evaluation of properties in pectic polymers. *Progress in Colloid and Polymer Science* 1998; 108:166–174.

51. Gemperline, P.J., Webber, L.D., Cox, F.O.. Raw Materials Testing Using Soft Independent Modeling of Class Analogy Analysis of Near-Infrared Reflectance Spectra. *Analytical Chemistry* 1989; 61(2):138–144.

52. Lee, S., Choi, H., Cha, K., Chung, H.. Random forest as a potential multivariate method for near-infrared (NIR) spectroscopic analysis of complex mixture samples: Gasoline and naphtha. *Microchemical Journal* 2013; 110:739–748.

53. Mevik, B.H., Segtnan, V.H., Næs, T.. Ensemble methods and partial least squares regression. *Journal of Chemometrics* 2004; 18(11):498–507.

54. Nawar, S., Mouazen, A.M.. Comparison between random forests, artificial neural networks and gradient boosted machines methods of on-line Vis-NIR spectroscopy measurements of soil total nitrogen and total carbon. *Sensors* 2017; 17(10):1–22.

55. DiFoggio, R.. Examination of Some Misconceptions About Near-Infrared Analysis. *Applied Spectroscopy* 1995; 49(1):67–75.

56. Næs, T., Isaksson, T., Fearn, T., Davies, T.. A User-Friendly Guide to Multivariate Calibration and Classification. Chichester, UK: NIR Publications; 2004.

57. Deshpande, A.A., Ramya, A., Vishweshwar, V., Deshpande, G.R., Roy, A.K.. Applications of Gage Reproducibility & Repeatability (GRR): Understanding and Quantifying the Effect of Variations from Different Sources on a Robust Process Development. *Organic Process Research & Development* 2014; 18(12):1614–1621.

Figure 1: Conceptual comparison of raw material characterization approaches qualitatively regarding uncertainty information gathered (and expected relative span) and time-effort it would require the manufacturer to implement the approach and its in-operation time expenditure. The box highlights the three approaches under quantitative comparison in this study.

Figure 2: Boxplots of samples grouped by lemon, lime and, orange for the measured response variables $\%Yield$, $\%DE_0$, IV_0 and, C^0_{pectin}.

Figure 3: C^0_{pectin} scatter plot. Samples colored by lemon, lime and, orange. Number labeled by supplier.

Figure 4: PC1 vs. PC2 biplot. Colored by lemon, lime and, orange, with the $\boldsymbol{u_{peel}}$ variables labelled.

Figure 5: PC1 vs. PC2 score plots. Colored by cluster classes obtained with Wards agglomerative algorithm (Mahalanobis distance). A) Includes all 69 samples B) Excludes the grey samples highlighted in A). The cluster analysis is repeated after the exclusion of the grey samples.

Figure 6: A) PC1 vs. PC2 score plot colored by lemon, lime and, orange, number labeled by the supplier. B) SNV and mean centered spectra. PC1+ indicates the samples highlighted in the box in A). C) Loading plots for the first two principal components.

Figure 7: Predicted vs measured $\%DE_0$ plot. Colored by lemon, lime and, orange. The dashed line is equivalent to 1:1 fit $\pm$RMSEP. The models for the other variables are not plotted but their performance is registered in Table 3

Figure 8: Average coefficient of variation (relative standard deviations) for the three approaches in comparison.

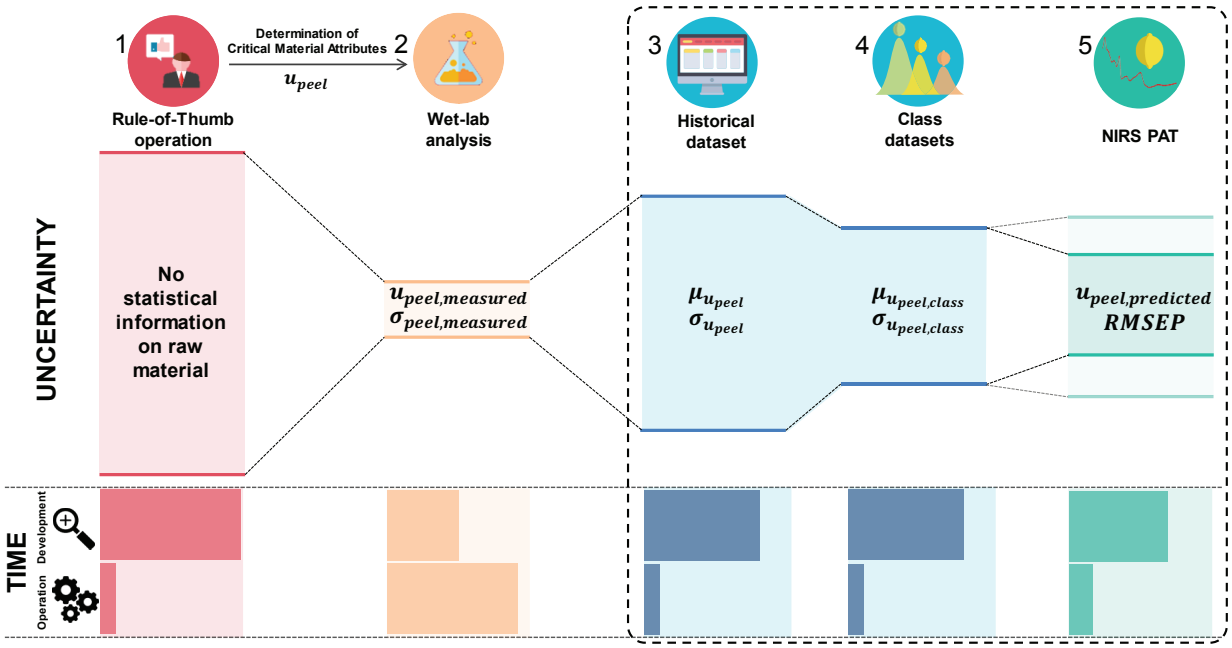

Figure 1
Double column size
Close to Raw Material Quality Assessment Approaches applied to Citrus Peel

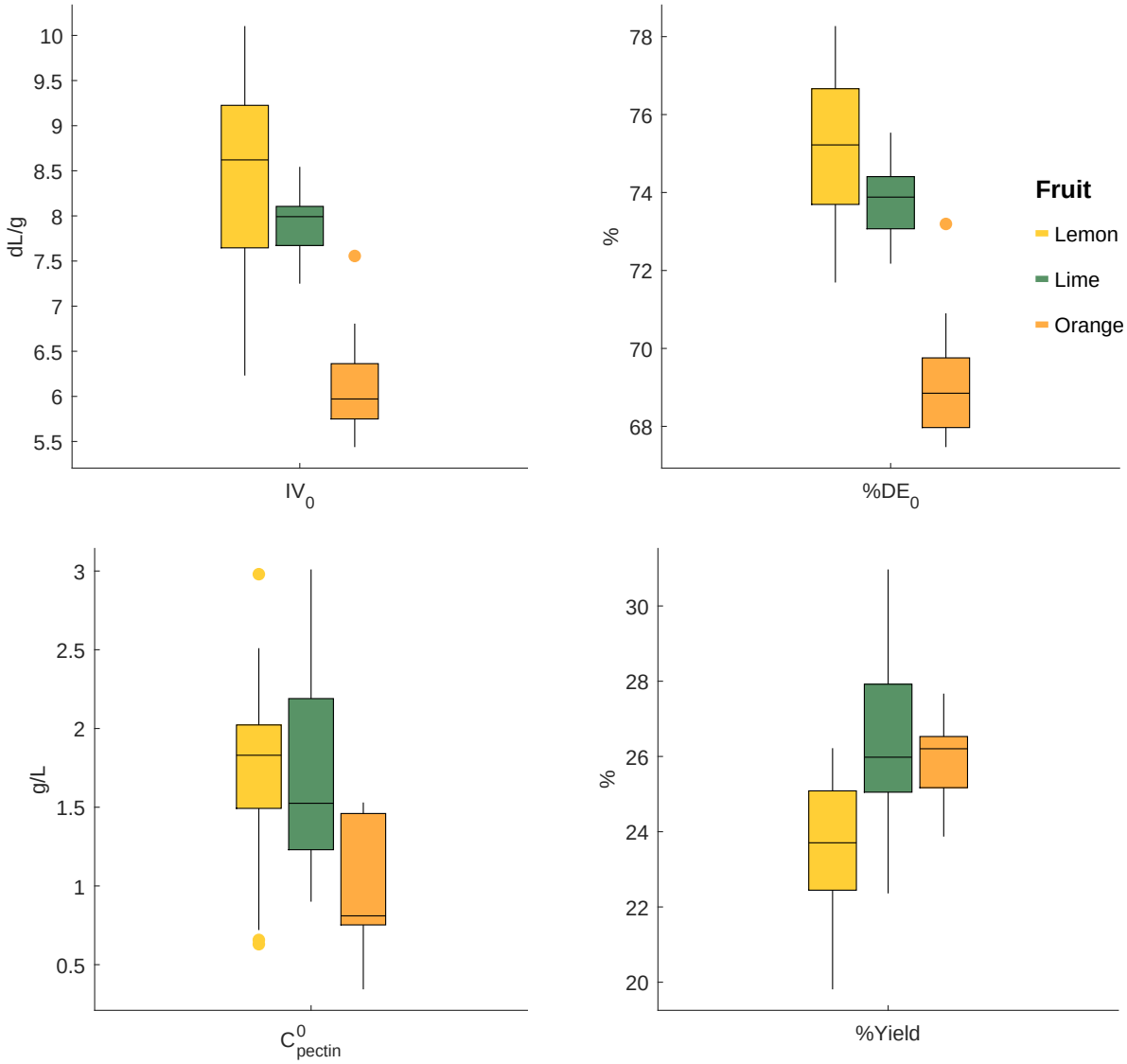

Figure 2
Double column width size
Close to Expert-knowledge classification

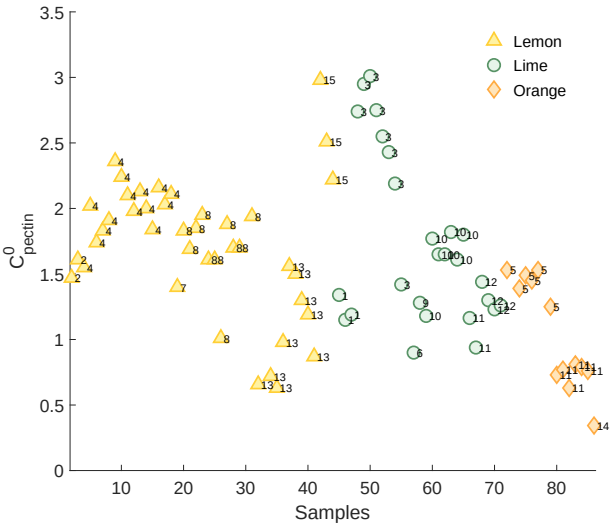

Figure 3
One column width size
Close to Expert-knowledge classification

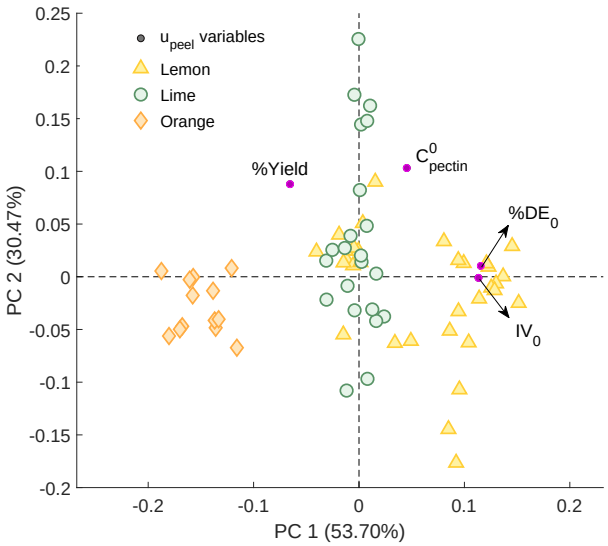

Figure 4
One column width size
Close to Unsupervised learning and cluster analysis

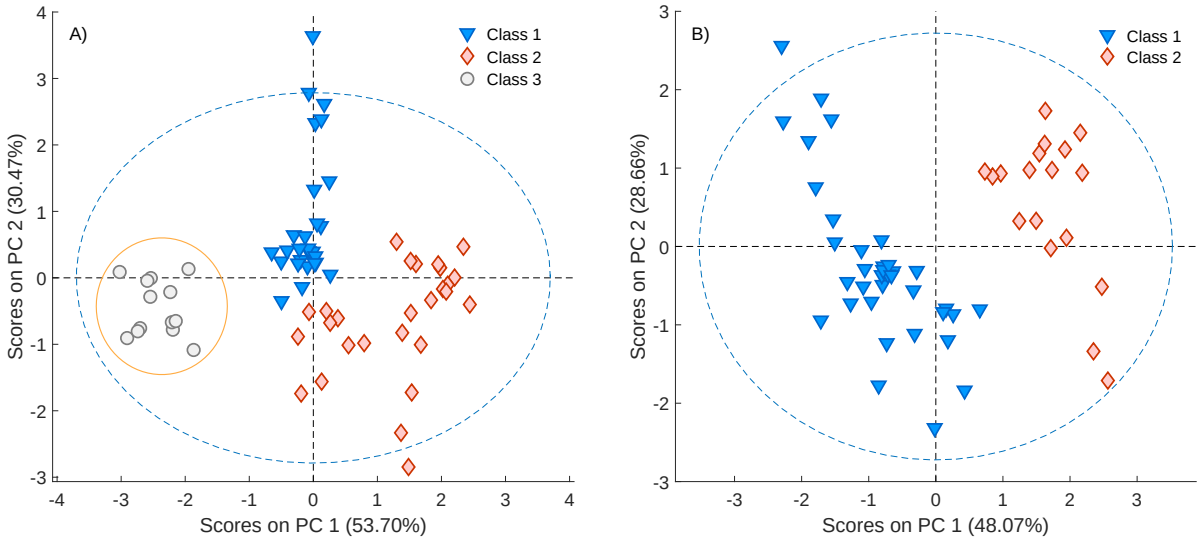

Figure 5
Double column width size
Close to Unsupervised learning and cluster analysis

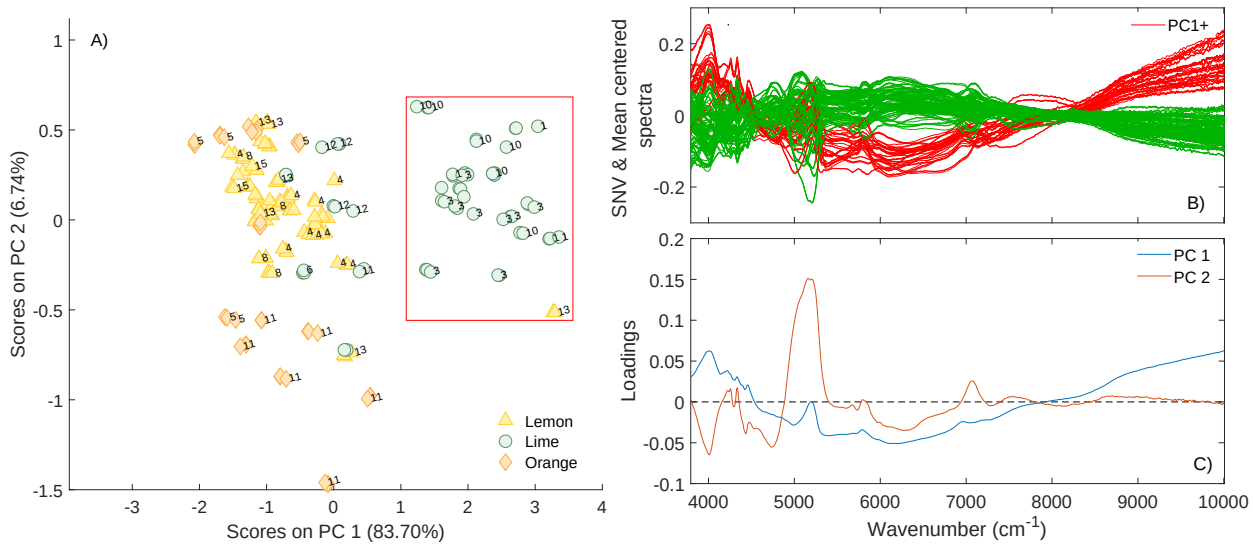

Figure 6
Double column width size
Close to Spectroscopic coupling

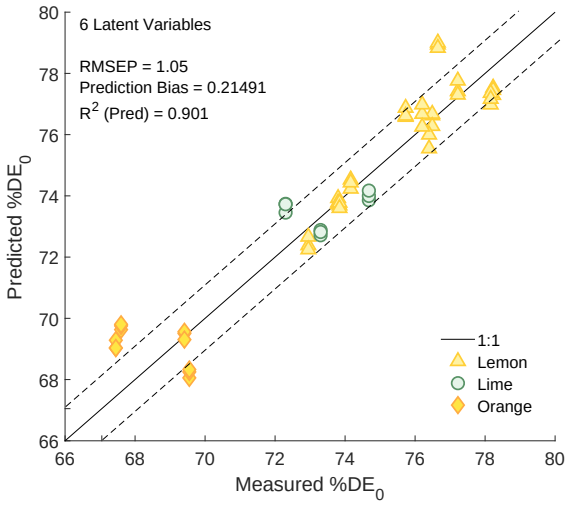

Figure 7
One column width size
Close to Critical material attributes prediction models

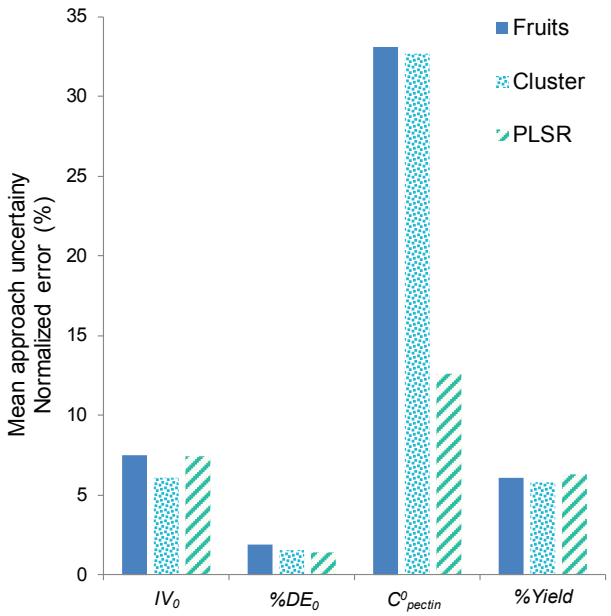

Figure 8
One column width size
Close to Quantitative statistical comparison

Table 1: Empirical measures of central tendency, spread, and asymmetry of the measured critical material variables. Full dataset included (n is the number o samples for each variable, μ the mean, σ the standard deviation and p10-p90 are the percentiles).

	$\%Yield$	$\%DE_0$	IV_0 (dL/g)	C^0_{pectin} (g/L)
n	68	63	68	80
μ	24.98	73.50	7.78	1.59
σ	2.04	2.81	1.14	0.59
min	19.81	67.47	5.44	0.34
max	30.97	78.27	10.10	3.01
p10	22.12	68.85	5.89	0.77
p25	23.69	72.18	7.24	1.23
p50	25.14	73.83	7.97	1.61
p75	26.22	75.53	8.60	1.95
p90	27.67	77.03	9.36	2.43
skewness	0.13	-0.52	-0.28	0.27
kurtosis	3.28	2.6	2.44	2.79

Close to Historical dataset statistics

Table 2: PLS-DA classification confusion matrix for a cross-validated set (keeping the spectral replicates together) for both scenarios of classes presented in Figure 5. Both models are built with full spectra and 10 latent variables.

		Actual Classes		
		Class 1	Class 2	Class 3
Predicted	Figure 5A (All Samples)			
	Class 1	61	18	0
	Class 2	16	56	0
	Class 3	0	0	38
	Figure 5B (Lemon samples)			
	Class 1	37	0	-
	Class 2	2	50	-

Close to Classification aiding tool

Table 3: Comparison of the three raw material quality assessment approaches

	Fruits			Cluster Classes			
	Lime	Lemon	Orange	1	2		PLS-R
IV_0							
μ	7.92	8.4	6.01	7.74	9.12	R^2	0.82
σ	0.31	0.98	0.41	0.48	0.47	RMSEP	0.58
$\%DE_0$							
μ	73.85	75.1	68.71	73.59	76.55	R^2	0.9
σ	1.00	1.98	1.15	1.04	1.11	RMSEP	1.05
C^0_{pectin}							
μ	1.72	1.72	1.11	1.81	1.69	R^2	0.84
σ	0.64	0.51	0.36	0.59	0.56	RMSEP	0.2
$\%Yield$							
μ	26.22	23.69	26.03	25.57	23.02	R^2	0.53
σ	2.07	1.57	0.96	2.00	1.37	RMSEP	1.57

Close to Quantitative statistical comparison